Artificial Intelligence Review **13:** 437–466, 1999.

Medical Data Mining on the Internet: Research on a Cancer Information System

ANDREA L. HOUSTON[1], HSINCHUN CHEN[1], SUSAN M. HUBBARD[2], BRUCE R. SCHATZ[3], TOBUN D. NG[1], ROBIN R. SEWELL[4] and KRISTIN M. TOLLE[1]

[1]*Management Information Systems Department, University of Arizona, Tucson, Arizona 85721*
(E-mail: ahouston@bpa.arizona.edu; hchen@bpa.arizona.edu; TNg@bpa. arizona.edu; ktolle@bpaosf.bpa.arizona, edu)
[2]*National Cancer Institute, Bethesda, MD 20852 (E-mail: su@icicb.nci.nih.gov)*
[3]*University of Illinois at Urbana-Champaign, Urbana, IL 61801*
(E-mail: schatz@csl.ncsa.uiuc.edu)
[4]*School of Library Science, University of Arizona, Tucson, Arizona 85721*
(E-mail: rrs@ai2.bpa.arizona.edu)

Abstract. This paper discusses several data mining algorithms and techniques that we have developed at the University of Arizona Artificial Intelligence Lab. We have implemented these algorithms and techniques into several prototypes, one of which focuses on medical information developed in cooperation with the National Cancer Institute (NCI) and the University of Illinois at Urbana-Champaign. We propose an architecture for medical knowledge information systems that will permit data mining across several medical information sources and discuss a suite of data mining tools that we are developing to assist NCI in improving public access to and use of their existing vast cancer information collections.

Keywords: CancerLit, concept spaces, data mining, Hopfield net, information retrieval, Kohonen net, medical knowledge, neural networks

1. Introduction

Through Executive Order No. 12864 President Clinton established in Advisory Council on the National Information Infrastructure (NII) in 1993 to identify appropriate government actions related to NII development. This initiative signalled a change in attitude toward availability of government information. The last five years have produced explosive growth in the amount of government information that is available to the general public through the Internet. Several government agencies, including the National Institutes of Health (NIH), National Science Foundation (NSF), National Library of Medicine (NLM), and National Cancer Institute (NCI) have

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1999		**00-00-1999 to 00-00-1999**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Medical Data Mining on the Internet: Research on a Cancer Information System	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Arizona ,Management Information Systems Department ,Tucson,AZ,85721**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This paper discusses several data mining algorithms and techniques that we have developed at the University of Arizona Artificial Intelligence Lab.We have implemented these algorithms and techniques into several prototypes, one of which focuses on medical information developed in cooperation with the National Cancer Institute (NCI) and the University of Illinois at Urbana-Champaign.We propose an architecture for medical knowledge information systems that will permit data mining across several medical information sources and discuss a suite of data mining tools that we are developing to assist NCI in improving public access to and use of their existing vast cancer information collections.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**30**	

developed web sites that serve as gateways to their extensive collections of information. Other agencies such as National Institute of Standards and Technology (NIST), DARPA and NASA are funding initiatives to help government agencies and businesses create useful knowledge from their wealth of raw data. This means that vast government collections of information, once available only from government document collections at libraries or government agencies, are now publicly accessible over the Internet. Users no longer have to request information through the mail, trusting the government agencies to summarize and categories it in a manner that is useful to the specific user, or travel to government agencies and request access to their collections. Information that was once available only in paper format is now available digitally, 24 hours a day, 7 days a week.

The new challenge for government organizations is to help interested individuals utilize government information in timely and meaningful ways. Vast collections of raw data are not in themselves useful. To be meaningful, data must be analyzed and converted into information, or even better, into knowledge. The amount of information available is staggering (CancerLit, a relatively small and narrow collection of biomedical information specific to cancer contains over one million documents). Traditional methods of data analysis utilizing human beings as pattern detectors and data analysts cannot possibly cope with such a large volume of information. Even more technologically sophisticated approaches such as spread-sheet analysis and ad-hoc queries cannot be scaled up to deal with tremendous amounts of raw information. However, data mining tools and knowledge discovery in databases (KDD) techniques are very promising approaches to help agencies provide information in meaningful ways.

2. Data Mining and Knowledge Discovery

Finding useful patterns or meaning in raw data has variously been called KDD (knowledge discovery in databases), data mining, knowledge discovery, knowledge extraction, information discovery, information harvesting, data archeology and data pattern processing (Fayyad et al. 1996a). In this paper, we will use Fayyad et al.'s (1996b) definitions of knowledge discovery and data mining. Knowledge discovery is the "non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data." Data mining, one of the steps in the process of knowledge discovery, "consists of applying data analysis and discovery (learning) algorithms that produce a particular enumeration of patterns (or models) over the data." Data mining is typically a bottom-up knowledge engineering strategy (Khosla and Dillon 1997). Knowledge discovery involves the additional steps of

target data set selection, data preprocessing, and data reduction (reducing the number of variables), which occur prior to data mining. It also involves the additional steps of information interpretation and consolidation of the information extracted during the data mining process.

Agrawal et al. (1993) proposed three basic classes of data mining problems (which we have further divided in to four classes). The nature of the data mining problem often suggests a certain class of data mining technique or method to be an appropriate solution.

- Classification – This problem involves the need to find rules that can partition the data into disjoint groups. Often classification involves supervised data mining tools in which the user is heavily involved in the definition of the different groups and the specification of the rules that can be used to determine to which group a data item belongs. Examples of such tools include decision trees and rule-based techniques. Example-based data mining methods such as nearest neighbor classification, regression algorithms and case-based reasoning are also examples of solutions to data classification problems (Fayyad et al. 1996a). For health care professionals, this type of data mining technique would be important in diagnostic and treatment assistance decision making. In our tools we use an automated classification technique that is based on Salton's vector representation and statistical analysis of documents. One of our tools, the concept space, uses a purely statistically based method to classify or index documents automatically. Another tool, the noun phrasing tool, syntactically parses documents by extracting valid noun phrases which are then statistically analyzed when classifying and indexing documents.
- Clustering – While Agrawal et al. (1993) consider clustering and classification to be in the same class of data mining problems, other researchers (for example Decker and Focardi 1995; Fayyad et al. 1996a; Holsheimer and Siebes 1994) consider clustering a separate class. This is because clustering methods allow data mining algorithms to determine groups automatically (or in an unsupervised manner), using actual data rather than classification rules imposed externally. Clustering techniques are frequently used to discover structure or similarities in data. Typically, clustering techniques are iterative in nature with a series of partitioning steps followed by a series of evaluation or optimization (of the partitions) steps, followed perhaps by a repartitioning and re-evaluation series of iterations. Feed-forward neural networks, adaptive spline methods and projection pursuit regression are all examples of this type of data mining solution. In the health care profession, this type of problem is especially interesting to researchers and health care insurers

or providers (i.e., HMOs and medical insurance companies) trying to discover information about a drug, a treatment or a disease. In our tools, clustering is accomplished by co-occurrence analysis, and the use of two different neural nets (the Hopfield net in the concept space and noun phrasing tools and the Kohonen net in the self-organizing map tools).

- Association – The association data mining problem involves finding all of the rules (or at least a critical subset of rules) for which a particular data attribute is either a consequence or an antecedent. This type of problem is very common in marketing data mining problems but is also of interest to health care professionals who are looking for relationships between diseases and life-styles or demographics or between survival rates and treatments, for example. Association problems are similar to rule-based methods, but in addition they typically have confidence factors associated with each rule. For this reason, an example of such a technique is a probabilistic graphical dependency model (with a Bayesian component). Often association-type data mining techniques are employed to help strengthen arguments concerning whether or not to include or eliminate candidate rules from a knowledge model.
- Sequences – This type of data mining problem involves ordered data, most commonly time sequence or temporal data. Stock market analysis is the most frequently used example. For health care professionals, disease progression and treatment success are two examples of medical information problems where a sequence-based data mining algorithm could be useful.

If data mining were simple, the world's information management problems would have been solved long ago. Unfortunately a good data mining technique must cope with a series of extremely difficult challenges. Some of these are: high dimensionality (very large number of attributes to compare and explore for relationships); missing, incomplete and noisy (or inaccurate) data; overfitting (this is a particular problems for clustering or automatic techniques); extremely complex relationships between variables that simplistic techniques cannot detect; integration between data sources and data types; volatility of some of the data and knowledge; assessing the statistical significance of the patterns detected; the impact on the results that data preprocessing has; HCI issues, such as the visualization issue that plagues result interpretation and the lack of human ability to understand some of the complex patterns that a computer can detect; privacy issues (especially relevant to medical information); and the meaningfulness of the patterns detected (i.e., are they interesting and useful or merely obvious) (Uthurusamy 1996).

Obviously our tools cannot address all of these challenges, but we believe that our techniques are useful in addressing: 1) high dimensionality; 2) overfitting; 3) missing, incomplete or noisy data; 4) visualization issues; and 5) privacy. The automatic indexing portion of the concept space tool and the part-of-speech tagger represent our attempts to reduce the high dimensionality of a full-text document collection data mining task. Each technique, when combined with the statistical analysis and co-occurrence analysis, identifies the 1,000–5,000 most relevant document descriptors in the collection as opposed to every existing document descriptor in the collection. Overfitting is more of a challenge, especially for the neural net components, as they can be over-trained, which results in overfitting. We control this by manipulating several parameters in both the clustering algorithms and the neural net training. For the CancerLit collection, we ran several test runs on smaller data sets with a variety of parameter settings, and chose the most promising parameters to use on the large collection. Years of experimenting with a variety of collections has yielded a set of default parameters and provided experience in what other combinations to try on new collections.

We have good evidence from electronic brain-storming session data that the neural net components of our tools are particularly good at coping with missing, incomplete or noisy data. The current medical text testbed (CancerLit), doesn't suffer from this problem because it is based on published articles from professional journals, and conferences. Two of our tools, the graphical concept space and the dynamic SOM, were specifically designed to help address the problem of visualization and we are continually investigating other visualization methods looking for improvements. Privacy is an issue in some of the testbed collections that we have worked with (for example COPLINK, a project with the National Institute of Justice and the Tucson Police Department), but it is not an issue with the CancerLit collection which is available to the public and does not contain any personal medical history or other personally sensitive data. However, the tools can be modified to increase security.

Data mining on the Internet also presents the problem of mining unstructured data. We encountered this problem when doing usability studies of our data mining algorithms on a part of the Internet (the Entertainment sub-directory of Yahoo!) (Chen et al. 1998a). Homepages have the same unstructured data problems common to any type of free-text data (electronic brain storming sessions, LotusNotes databases, or collections of e-mail, for example). In addition, many personal homepages lack the coherence or unifying theme of other free-text data collections. In most textual collections, each document has a major subject or theme (i.e., it is about a given topic or set of topics). The unifying theme in a personal homepage is the interests

of the author, which often are eclectic and have nothing in common except the author's interest. This can cause data mining algorithms searching for relationships in unstructured data to come up with nonsensical categories or associations.

We chose medical literature as our application area for this prototype. Specifically, we selected NCI's CancerLit collection because, while it has many of the data mining challenges mentioned above (e..g, high dimensionality, complex relationships, HCI – human computer interaction issues, privacy issue and the meaningfulness of detected patterns), it has a structure and each document has a major topic or theme, making it easier than Internet personal homepages to apply our data mining tools. Each document has a set of tokenized fields that contain free-text information. Examples are: author, author address, publication information (i.e., where each document was published), title, MeSH indexing terms (manually assigned by NLM indexers to each document), and abstract. In essence, the CancerLit collection can be treated as a large semi-structured textual database.

3. The National Cancer Institute (NCI)

The National Cancer Institute has the following stated mission: "The National Cancer Institute coordinates the National Cancer Program, which conducts and supports research, training, *health information collection and dissemination*, and other programs with respect to the cause, diagnosis, prevention, and treatment of cancer, rehabilitation from cancer, and the continuing care of cancer patients and the families of cancer patients" (Quote from director's homepage).

NCI (the National Cancer Institute) is responsible for managing an immense collection of cancer-related information. Part of that information management responsibility involves finding innovative ways to share information in as timely, efficient, and intuitive manner as possible. NCI has therefore instituted a series of small information-sharing initiatives which are publicly available on-line through various links to their World Wide Web (WWW) pages. NCI also shares its digitized collections in a variety of formats (including CD-ROM) as testbeds for data mining investigations. Some of NCI's on-line initiatives involving cancer information include:

- **CancerNet** (http://www.nci.nih.gov) – provides information about cancer, including state-of-the-art information on cancer screening, prevention, treatment and supportive care, and summaries of clinical trials.
- **CancerNet for Patients and the Public** – includes access to PDQ (Physician Data Query) and related information on treatments; detec-

tion, prevention and genetics information; supportive care information; clinical trial information; a directory of genetic counselors; a multimedia breast cancer resource; a Kid's HomePage; and information on the Cancer Information Service, a toll-free information service. (http://cancernet.nci.nih.gov/patient.htm);

- **CancerNet for Health Professionals** – includes access to PDQ and related information on: treatments; screening, prevention and genetics; supportive care and advocacy issues; clinical trials; a directory of genetic counselors; CancerLit topic searches; cancer statistics; and the *Journal of the National Cancer Institute*. (http://wwwicic.nci.nih.gov/health.htm);
- **CancerNet for Basic Researchers** – includes access to a cooperative breast cancer tissue research database; a breast cancer specimen and data information system; an AIDS malignancy bank database; a cooperative human tissue network; a cooperative family registry for breast cancer studies; cancer statistics (SEER); NCI Information Associates Program and the *Journal of the National Cancer Institute*. (http://wwwicic.nci.nih.gov/research.htm); and
- **CancerLit** – a comprehensive archival file of more than one million bibliographic records (most with abstracts) describing 30 years of cancer research published in biomedical journals, proceedings of scientific meetings, books, technical reports, and other documents. CancerLit increases by more than 70,000 abstracts each year. The database is updated monthly to provide a comprehensive, up-to-date resource of published cancer research results. Preformulated searches for over 80 clinical topics are updated and available on the Web site. The complete CancerLit database is now searchable on the Web at: (http://cnetdb.nci.nih.gov/canlit/canlit.htm). It is also available through the National Library of Medicine and through a variety of commercial database vendors (as a CD-ROM product). (http://wwwicic.nci.nih.gov/canlit/canlit.htm).

NCI's International Cancer Information Center (ICIC) clearly considers that it is essential for the cancer information that it manages to be easily accessible to all levels of medical information users from the very naive to the extremely expert. "Other novel channels of information distribution will be explored to bring cancer information to those who require it, whether health professionals, patients, or policy makers. Appropriate choice cannot be made unless the full range of options is available to these decision makers" (Hubbard et al. 1995). The deployment of PDQ – Physician Data Query (a current peer-reviewed synthesis of state-of-the art clinical information on cancer) is a good example. The PDQ knowledge source is available in a

variety of formats: local access via purchased CD-ROMs for PC use; CancerFax (a fax-on-demand system); CancerNet (an Internet service described above); CancerMail (an e-mail application); and NCI's Associates Program (a membership program that provides direct access to scientific information services).

ICIC (International Cancer Information Center) is constantly investigating emerging technologies to find ways to improve the content, timeliness of dissemination, accessibility and use of cancer information in general and PDQ information in particular. A good example is the incorporation of Mosaic, a client-server based application (developed by the National Center for Supercomputing Applications – NCSA at the University of Illinois) that is platform independent. The use of Mosaic allows the delivery of interactive multimedia information to users and permits the dissemination of cancer-related graphical images, sound and full-motion video in addition to the existing text-based cancer information already available.

ICIC's goal is to identify future systems that will assist users in finding pertinent information and in determining how data pertain to specific clinical situations (Hubbard et al. 1995). As part of this goal, ICIC (International Cancer Information Center) must have a vision of how it wants the cancer information disseminated and an architecture that will support the dissemination process. This architecture must be scalable, because the amount of cancer information available is vast and continues to increase at an incredible rate. The volume of available cancer information is too overwhelming to be accessed without some kind of organization which allows users to extract information of interest to them in manageable units (i.e., data mining). The volume of information involved has encouraged ICIC to seek automatic (as opposed to manual) solutions to the challenge of indexing and categorizing its information collections.

Another challenge to medical information systems is the diversity of users, especially with respect to their levels of subject area expertise, their knowledge and understanding of the information sources (in particular knowing how to query and navigate the information sources to extract useful information), and their information usage requirements. In addition, the cancer information managed by ICIC comes from a variety of sources in a variety of formats, only some of which are pre-indexed. This means that any NCI (National Cancer Institute) information system architecture must be easy-to-use (to permit accessibility by users of all levels of expertise and source familiarity) and flexible (to accommodate the diversity of information sources and information use), in addition to being scalable.

"Any medical information system must be designed around an architecture that provides security, scalability, modularity, manageability, and cost-

effectiveness" (Varnado 1995). At the University of Arizona, in conjunction with the University of Illinois at Urbana-Champaign, we have been working with the (International Cancer Information Center (ICIC) at the National Cancer Institute (NCI) to develop such a medical information retrieval architecture. We are particularly interested in an architecture that will support a variety of easy-to-use data mining tools and one that will support Internet access to the CancerLit collection using our data mining tools. Figure 1 is an illustration of that architecture. It is based on the concept of multiple integrated information management and data mining tools that can address these scalability, modularity, manageability, diversity, efficiency, timeliness and cost-effectiveness challenges/requirements.

The medical information retrieval system architecture consists of a suite of tools (described in more detail in the next section) which incorporate a number of data mining techniques focusing on categorization (which occurs in the automatic indexing of textual and image data sources step) and clustering (which occurs in the data inductive learning and data analysis step and primarily takes advantage of statistical co-occurrence analysis, and the Hopfield net and Kohonen net algorithms, although we are exploring other clustering techniques such as Ward's algorithm and Multi-dimensional Scaling). The result is a set of concept space which we link to other pre-existing indexing and summarization sources (i.e., MeSH terms, and the Unified Medical Language System – UMLS Metathesaurus). Users can explore this set of concept spaces using self-organizing maps (i.e., Kohonen), and spreading activation techniques (i.e., the Hopfield net algorithm). The concept spaces are linked via commonly held vocabulary and document descriptors, which serve as bridges or gateways between them. This allows information seekers to explore a variety of concept spaces using a tool that they are comfortable with that will suit their specific information searching needs. Some tools are better suited to narrow and specific searches, while others are better suited to browsing. Currently, the medical system provides individual access to each tool from a single web-page source. However, the Arizona Artificial Intelligence Lab has developed a Geographic Information System (GIS) system as a DARPA prototype that fully integrates different knowledge sources and data types into one completely integrated query system.

4. A Suite of Tools for Improving Access to NCI Information

The University of Arizona AI Lab is working with the University of Illinois at Urbana-Champaign and the National Cancer Institute (NCI) to develop a suite of data mining tools as part of the implementation of the medical information retrieval system architecture mentioned above. Our goal is to

help improve access to and analysis of NCI's cancer information. We have been using the CancerLit collection as our testbed for tool development and tool usability studies. Some of the tools are based on technology developed in other research projects, and some have been developed specifically for the CancerLit collection. The following sections briefly describe these tools and the type of user and medical information needs for which we believe they will be useful. Most of this work is in the development stage and we have performed only pilot tests to assist in the design and implementation process. All of the tools were designed to be accessed via the World Wide Web (WWW).

4.1 *Concept spaces*

A concept space is an automatically generated thesaurus that is based on terms extracted from documents (see Chen et al. 1996a; Chen et al. 1993 for details). Our CancerLit concept space identified terms from three different sources: proper names from the author field; MeSH (Medical Subject Headings) terms that the National Library of Medicine (NLM) indexers had assigned to the document from the keyword field; and descriptor terms (called automatic indexing terms) created from the free-text in both the title and abstract fields during the automatic indexing and clustering stages. This process is a syntactical Salton-based technique (Salton 1989) which represents documents in a collection as a set of multiple descriptor term vectors. The underlying algorithms include: a statistical co-occurrence algorithm developed by Chen and Lynch (1992) to identify document descriptor terms (which can be phrases that contain between one and five words), a threshold value to limit document descriptor terms to the more commonly occurring ones, an asymmetric weighting scheme to reward more specific terms (presumed to be more "interesting" or descriptive), and a Hopfield net algorithm (Hopfield 1982; Tank and Hopfield 1987) to identify relationships between the document descriptors.

The CancerLit concept space was created using the following process:

– **Document collection identification:** The first task, obviously, is to identify the document collection of interest. The only restrictions are that the collection must be digitized and that each document be delimited in some way (so that the analysis program can cleanly identify where one document ends and the next one begins). For collections where there are specific items of interest (for example in the medical literature collection: author, MeSH indexing terms, title, abstract and document source – i.e., journal name) each item must be distinguishable from the rest of the text. We selected CancerLit as the document collection for this research.

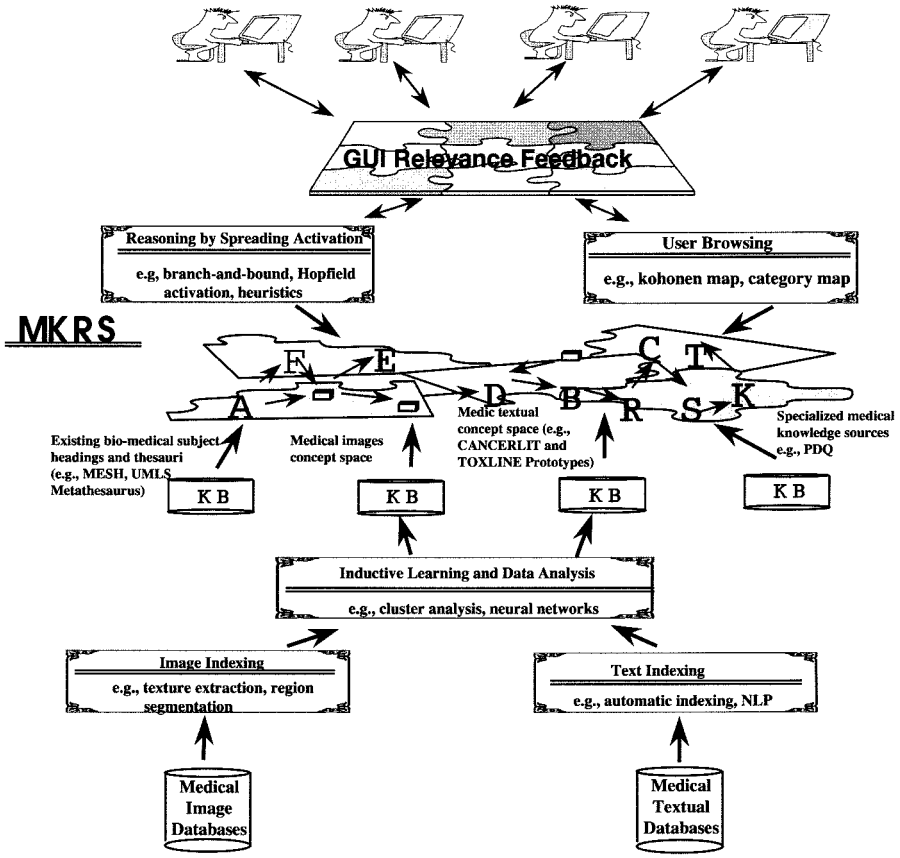

Figure 1. Medical information retrieval system architecture.

– **Objective filtering and automatic indexing:** The purpose of this step is to *automatically* identify each document's content. The first process in this step is *object filtering* which involves the identification of document descriptors in each document that match known knowledge domain vocabulary, ontology or standard indexing terms. In the CancerLit collection testbed the documents have already been indexed by MeSH (Medical Subject Heading) terms. In other document testbeds, this information had to be developed and a pattern matching program run against each document.

The next process uses a Salton-based automatic indexing technique (Salton 1989) (which typically includes dictionary look-up, stop-wording, word stemming, and term-phrase formation) that identifies document subject descriptors. Each word in the document is checked against a stop-word list which eliminates non-content bearing words

(e.g., “the”, “a”, “on”, “in”). Initially we used a generic stop-word list that was borrowed from other information retrieval research, but discovered that a high quality stop-word list is really collection and knowledge domain dependent. For example, removing the word *in* from medical literature is actually detrimental as there are several important phrases (i.e., *in vivo*, *in vitro*) where *in* is critical to the correct construction of the phrase. If a word is on the stop-word list, it is removed, otherwise it is kept as part of the analysis.

In Salton’s research, the next step would involve a stemming algorithm which reduces each word to its major word stem (for example, plural forms are reduced to the singular form and verb tenses are all changed to present tense). Our research with scientific documents suggests that, except in the reduction of plurals, this is not useful so we have removed the step altogether.

Next a term-phrase (or document descriptor) formation step that looks at combining adjacent words to form phrases is performed. Initially, we restricted the term-phrase formation to one, two and three word phrases, but for medical research, we discovered that four and five word phrases are also appropriate. In the noun phraser research we have used a part-of-speech tagger in combination with some simple syntax rules to help identify multiple word phrases. Once document descriptors are formed, a statistical analysis program calculates collection frequency, document frequency and inverse document frequency (which helps to identify the specificity of the descriptor). Only descriptors that meet a certain set of threshold requirements are kept as valid document descriptors.

We have done several term evaluation experiments, trying to determine the quality of the document descriptors identified by the object filtering and automatic indexing techniques. The strength of Salton’s technique is that the document descriptors identified actually come from the physical text of the document, i.e., they are the authors’ actual words. The enormous advantage that such a technique has over standardized indexing vocabulary terms, especially in scientific literature, is that the most up-to-date terminology and commonly used abbreviations and acronyms are always included as potential candidates for document descriptors. One of the biggest complaints about standard indexing terms is that they always lag behind current terminology. On the other hand, manual indexing by human beings frequently adds value to a document descriptor by choosing terms that may not actually occur in the document but are very relevant to the major theme or topic of the document (generalizations and summarizations are good examples). A combined approach of object filtering, automatic indexing and using any already

provided human indexing terms as document descriptors appears to be the best approach to identifying document descriptors.

– **Co-occurrence analysis:** The importance of each desciptor (term) in representing document content varies. Using term frequency and inverse document frequency, the cluster analysis step assigns weights to each document term to represent term importance. Term frequency indicates how often a particular term occurs in the *entire* collection. Inverse document frequency (indicating term specificity) allows terms to have different strengths (importance) based on specificity. In typical collections, a term can be a one-, two-, or three-word phrase (up to five-word phrase in medical and very detailed scientific collections). Figure 2 describes our frequency computation. Usually terms identified from the title of a document, the entire author name (normalized to last name, first name), and other terms identified in the object filter step are more descriptive than terms identified from other parts of the document. Therefore, these terms are assigned heavier weights than other terms (i.e., rewarded). Multiple-word terms are also assigned heavier weights than single-word terms because multiple-word terms usually convey more precise semantic meaning than single-word terms. Eventually we hope to incorporate even more intelligence in our weighting schemes, for example, by giving a higher weight to certain positions in a sentence (i.e., the sentence object or subject), in a paragraph (theoretically the leading and ending sentences of a paragraph are more important), and in the document (for example, terms found in the conclusion section of a journal paper might be more important).

  Cluster analysis is then used to convert raw data indexes and weights into a matrix indicating term similarity/dissimilarity using a distance computation based on Chen and Lynch's *asymmetric* "Cluster Function" (Chen and Lynch 1992), which represents term association better than the cosine function (see Figure 3 for more detail). A net-like concept space of terms and weighted relationships is then created, using the cluster function.

– **Associate retrieval:** The Hopfield net (Hopfield 1982) was introduced as a neural network that can be used as a content-addressable memory. Knowledge and information can be stored in single-layered, interconnected neurons (nodes) and weighted synapses (links) and can be retrieved based on the Hopfield network's *parallel relaxation* and *convergence* methods. The Hopfield net has been used successfully in such applications as image classification, character recognition, and robotics (Knight 1990; Tank and Hopfield 1987) and was first adopted for *concept-based* information retrieval in (Chen et al. 1993).

In the AI lab's implementation, each term (identified in the previous steps) in the networklike theasurus is treated as a neuron and the calculated asymmetric weight between any two terms is used as the unidirectional, weighted connection between neurons. Using user-supplied terms as input patterns, the Hopfield algorithm activates their neighbors (i.e., strongly associated terms), combines weights from all associated neighors (by adding collective association strengths), and repeats this process until convergence. During the process, the algorithm causes a *damping effect*, in which terms farther away from the initial terms receive gradually decreasing activation weights and activation eventually "dies out." This phenomenon is consistent with the human memory *spreading activation* process.

Each document collection's Hopfield net is created as follows:

1. *Initialization of the network with automatic indexing terms* – The document descriptors identified by the object filter and automatic indexing steps are used to initialize the nodes (neurons) in the network and the links are assigned a random weight value.
2. *Iterative activation and weight computation* – Next, the net is trained using the document vectors and co-occurrence weights calculated in previous steps. The formulas in Figure 4 show the parallel relaxation property of the Hopfield net. At each iteration, nodes in the concept space are activated in parallel and activated values from different sources are combined for each individual node. Neighboring nodes are traversed in order until the activation levels of nodes on the network gradually "die out" and the network reaches a stable state (convergence). The weight computation scheme ($net_j = \Sigma_{i=0}^{n-1} t_{ij} u_i(t)$) is unique to the Hopfield net algorithm. Each newly activated node computes its new weight based on the summation of the products of its neighboring nodes' weights and the similarity of its predecessor node to itself.
3. *Convergence condition* – The above process is repeated until there is no significant change in terms of output between two iterations, which is accomplished by checking the following formula:

$$\sum_{j=0}^{n-1} |u_j(t+1) - u_j(t)| \leq \varepsilon$$

where ε is the maximum allowable error (used to indicate whether there is a significant difference between two iterations). Once the network converges, the output represents the set of terms most relevant to the starting input terms.

The combined weight of term j in document i, d_{ij} is computed as follows:

$$d_{ij} = tf_{ij} \times \log(\frac{N}{df_j} \times w_j)$$

where N represents the total number of documents in the collection, tf_{ij} represents the number of occurrences of term j in document i, w_j represents the number of words in descriptor T_j, and df_j, represents the number of documents in a collection of n documents in which term j occurs. Multiple-word terms are assigned heavier weights as they usually convey more precise semantic meaning.

Figure 2. Frequency computation.

Previous research demonstrated that this process can classify information at least as well as humans and in dramatically less time (5 minutes vs. 1 hour) (Chen et al. 1996a; Chen et al. 1998b), that it is scalable and useful in classifying a very large eclectic collection (part of the World Wide Web) (Chen et al. 1998a), and that it can be used for vocabulary switching between two concept spaces (Chen et al., submitted?).

The main advantage of the concept space approach is that it uses terms actually contained in the document (a bottom-up approach) *and* MeSH (Medical Subject Headings) terms that have been assigned to the document by medical indexers from the National Library of Medicine – NLM (a top-down approach that takes advantage of an existing classification system). Furthermore, it allows users to identify relationships that occur between descriptor terms, which can help narrow information retrieval or can highlight unknown relationships that actually are discussed in the documents in the collection. Figure 5 illustrates a user search using the CancerLit concept space looking for information related to *breast cancer*.

Previous research (Chen et al. 1998a; Houston et al. 1998) has indicated that concept spaces are ideal for refining a broad search topic into a more specific search topic and for discovering relationships between document descriptors. We believe that concept spaces will be most effective in helping novice users or users exploring in an area, domain or collection outside their own expertise to narrow and focus their searches. Concept spaces have also already been used to identify existing relationships between items (word phrases) in documents in a collection. In at least once instance involving police reports, highlighting such relationships has helped detectives solve a crime involving a new computer-theft gang. Currently we are exploring a variety of term-based weighting schemes to improve concept space information retrieval quality including improving the "interestingness" or meaningfulness of both the document descriptor terms themselves and the relationships identified among them.

$$\textit{Cluster Weight}(T_j, T_k) = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}} \times \textit{Weighting Factor}(T_k)$$

$$\textit{Cluster Weight}(T_k, T_j) = \frac{\sum_{i=1}^{n} d_{ikj}}{\sum_{i=1}^{n} d_{ik}} \times \textit{Weighting Factor}(T_j)$$

These two equations indicate the similarity weights from term T_j to term T_k (the first equation) and from term T_k to term T_j (the second equation). d_{ij} and d_{ik} are frequency computations from the previous step. d_{ijk} represents the combined weight of both descriptors T_j and T_k in document i defined as:

$$d_{ijk} = tf_{ijk} \times \log\left(\frac{N}{df_{jk}} \times w_j\right)$$

Co-occurrence analysis *penalizes* general terms using the following weights (similar to the *inverse document frequency* function), allowing the thesaurus to make more precise suggestions:

$$\textit{Weighting Factor}(T_k) = \frac{\log\frac{N}{df_k}}{\log N}$$

$$\textit{Weighting Factor}(T_j) = \frac{\log\frac{N}{df_j}}{\log N}$$

Figure 3. Cluster analysis computations.

$$u_i(t) = x_i, \quad 0 \le i \le n-1$$

$u_i(t)$ is the output of node i at time t. x_i (which has a value between 0 and 1) indicates the input pattern for node i.

$$u_j(t+1) = f_s\left[\sum_{i=0}^{n-1} t_{ij} u_i(t)\right], \quad 0 \le j \le n-1$$

where $\mu_j(t+1)$ is the activation value of term j at iteration $t+1$, t_{ij} is the co-occurrence weight from term i to term j, and f_s is the continous SIGMOID transformation function (which normalizes any input to a value between 0 and 1) as shown below (Dalton and Deshmane 1991; Knight 1990):

$$f_s(net_j) = \frac{1}{1 + exp\left[\frac{-(net_j - \theta_j)}{\theta_0}\right]}$$

where $net_j = \sum_{i=0}^{n-1} t_{ij} u_i(t)$, θ_j serves as a threshold or bias, and θ_0 is used to modify the shape of the SIGMOID function. (Chen and Ng (1995) provides more algorithmic detail.)

Figure 4. Hopfield net parallel relaxation formulas.

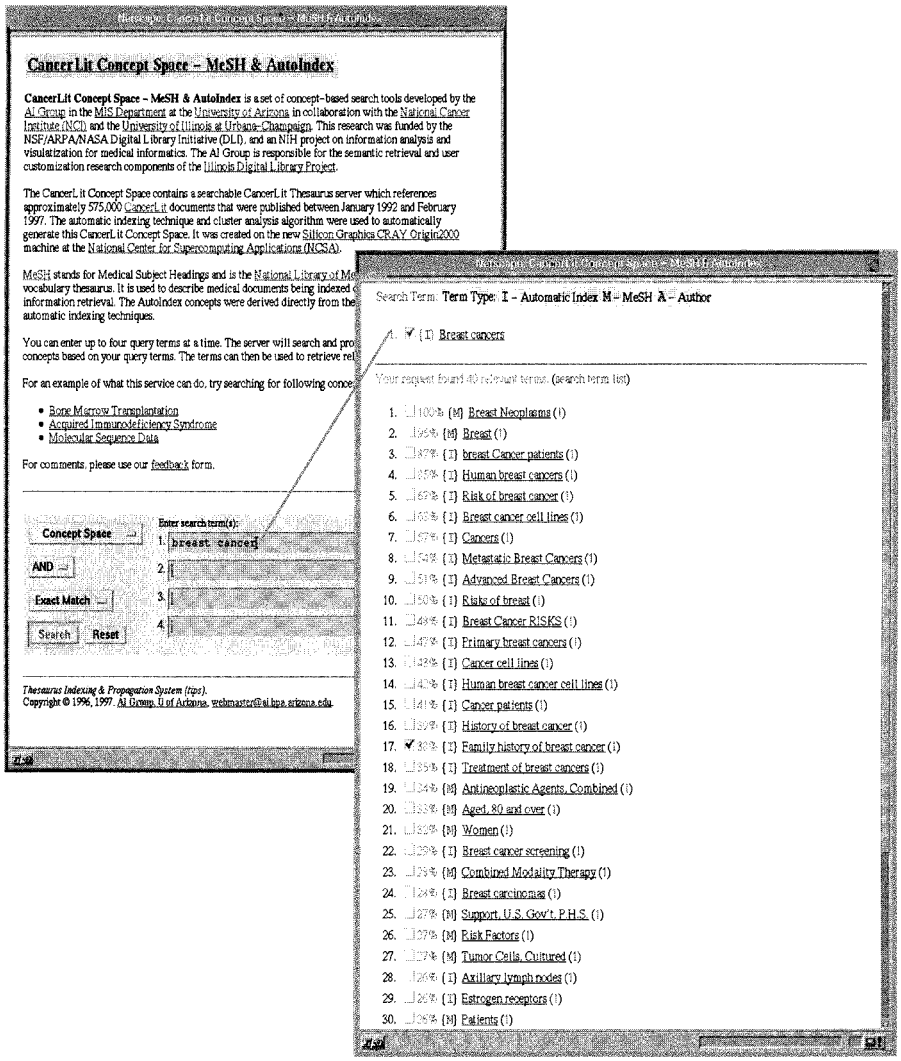

Figure 5. Automatic Indexing Session: User entered “breast cancer”. CancerLit Concept Space suggests terms related to “Breast cancers”.

4.2 *Arizona Noun Phraser*

The Arizona Noun Phraser research is investigating the potential of combining traditional keyword and syntactic approaches with a semantic approach to improve information retrieval quality. A common criticism of keyword based data mining techniques and searches is that single word terms lack an appropriate level of context for meaning evaluation. Incorporating a noun

phrase parser into the descriptor term identification phase of the algorithm would enable information systems to identify noun phrases and evaluate word meaning in the context of an entire noun phrase, potentially improving the precision and level of detail of information retrieval and the quality of the relationships identified. The Arizona Noun Phraser is based on the Brill tagger developed at the University of Pennsylvania (Brill 1993). Details can be found in Tolle (1997).

Once the document descriptor terms are identified by the Arizona Noun Phraser, they are processed by the clustering algorithms described above, creating in essence a second concept space. The major difference between the noun phraser and the automatic indexing concept spaces is that the automatic indexer concept space identifies document descriptor terms solely on the basis of statistics without regard to syntax, while the Arizona Noun Phraser initially identifies document descriptor terms on a syntactical basis. Both techniques use the same algorithms to evaluate the quality or usefulness of the document descriptor terms identified, and the same algorithms to identify relationships. Figures 6 illustrates a data mining exercise using the Arizona Noun Phraser in which a user is exploring document descriptor terms related to the term "breast cancer" and the documents that contain the relationships of interest.

This tool was developed specifically in response to a perceived need to quickly access medical information on a very narrowly defined, precise (or deep) topic. Pilot studies (Houston 1996; Houston 1998) indicate that it should be of value to experts or users with very narrow, detailed, and deep directed search and data mining requirements, such as Cancer Institute researchers or primary health care professionals using patient record information. We are currently experimenting with tuning the Arizona Noun Phraser by using stopwords, stop phrases, and adjusting term weights for various kinds of noun phrases (based on phrase length, noun phrase type, and source of noun phrase in the document) to improve the quality of the information retrieved and the relationships identified among the data.

Furthermore, we are looking at improving the quality of relationship identification in both tools (Arizona Noun Phraser and Concept Space) by using a lexicon to identify synonyms for the document descriptor terms so that obvious relationships can be ignored and more interesting and unusual ones identified. Allowing a user to bundle together document descriptor terms that are synonyms would allow a user to identify relationships among concepts, a larger unit or object than a simple document descriptor.

4.3 *Category maps*

Category maps are Kohonen-based self-organizing maps (SOMs) or neural nets that are associative in nature (see Chen et al. (1996b) for algorithmic

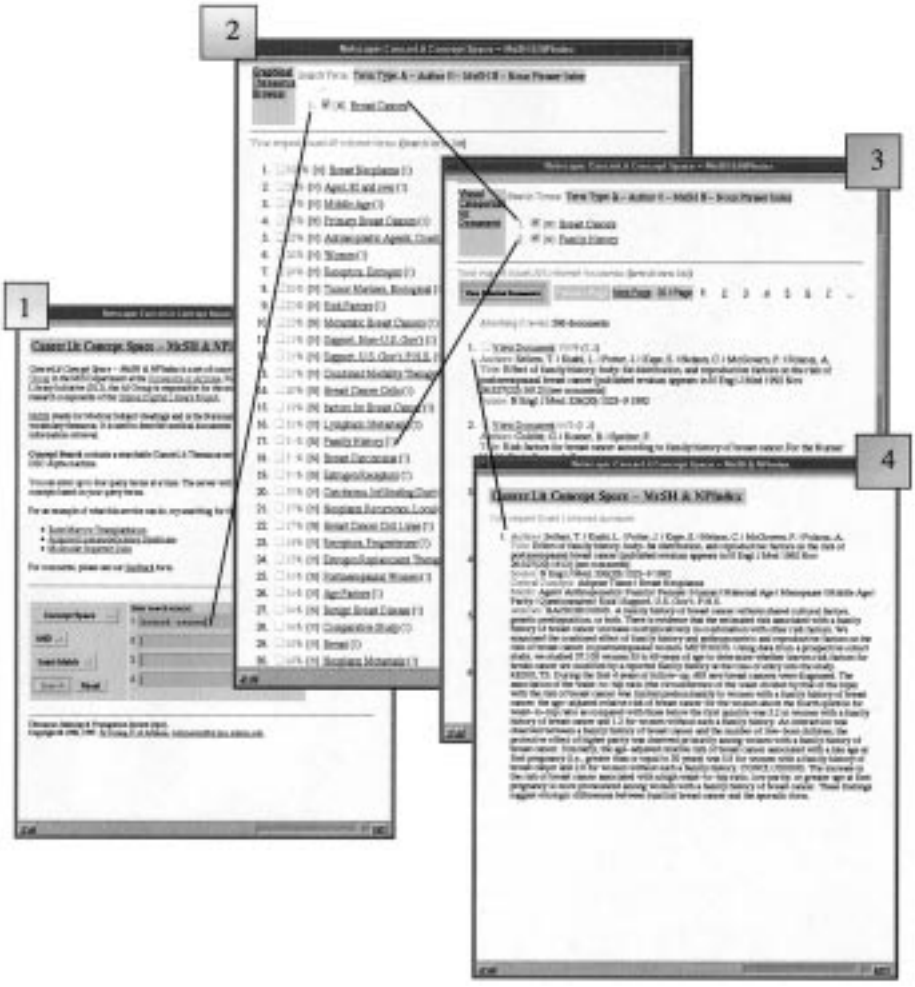

Figure 6. Arizona Noun Phraser Session: User entered “breast cancer” in (1). CancerLit Noun Phraser suggests related terms in (2). User selects “Family History”. System locates 506 documents related to “Breast cancers” and “Family History” in (3). User chooses to read first document in (4).

details) (Kohonen 1989; Kohonen 1995; Ritter and Kohonen 1989). The CancerLit map was created from document terms identified as part of the Concept Space automatic indexing process and then clustered using a modified Kohonen SOM algorithm. Our modification creates a multi-layered SOM algorithm, which permits unlimited layers of Kohonen maps (we refer to it as M-SOM). A sketch of the M-SOM algorithm is presented below:

1. **Initialize input nodes, output nodes, and connection weights:** Use the top (most frequently occurring) N terms (typically 1000 – 5000) from the

entire collection as the input vector and create a two-dimensional map (grid) of M output nodes (usually a 20-by-10 map of 200 nodes). Initialize weights from N input nodes to M output nodes to small random values.

2. **Present each document in order:** Represent each document by a vector of N terms and present to the system.
3. **Compute distances to all nodes:** Compute distance d_j between the input and each output node j using

$$d_j = \sum_{i=0}^{N-1} (x_i(t) - w_{ij}(t))^2$$

where $x_i(t)$ is the input to node i at time t and $w_{ij}(t)$ is the weight from input node i to output node j at time t.

4. **Select winning node j* and update weights to node j* and neighbors:** Select winning node j* as that output node with minimum d_j. Update weights for node j* and its neighbors to reduce their distances (between input nodes and output nodes). (See Kohonen (1989) and Lippmann (1987) for the algorithmic detail of neighborhood adjustment.)
5. **Label regions in map:** After the network is trained through repeated presentation of all document vectors (each vector is presented at least 5 times), submit unit input vectors of single terms to the trained network and assign the winning node the name of the input term. Neighboring nodes which contain the same name/term then form a concept/topic region (group). Similarly, submit each document vector as input to the trained network again and assign it to a particular node in the map. The resulting map thus represents regions of important terms/concepts (the more important a concept, the larger a region) and the assignment of documents to each region. Concept regions that are similar (conceptually) will also appear in the same neighborhood.
6. **Apply the above steps recursively for large regions:** For each map region which contains more than k (i.e., 100) documents, conduct a recursive procedure of generating another self-organizing map until each region contains no more than k documents.

We believe that, with 3–4 layers of self-organizing maps and a simple subject category browsing interface, we can partition the CancerLit collection into meaningful and manageable sizes that represent the major topics and relationships in the collection. In this implementation, the color of a map region has no significance other than to help differentiate each region from its neighbors. In other testbed collections (including electronic brain storming sessions and a commercial repair record database), color was used to indicate the recency of the theme or repair problem. Figure 7 illustrates a user's exploration of three levels of the map.

Our SOM is based on the same algorithms as the WEBSOM project (Honkela et al. 1998; Honkela et al. 1996a; Honkela et al. 1996b; Kaski 1996) (see (http://websom.hut.fi/websom/). The major differences between the two projects are: 1) the document encoding schemes are different; 2) the scale is different (as are the collections); 3) the visualization/user interface is very different; and 4) our implementation is a multiple-layer map.

Previous research (Chen et al. 1998a; Houston 1996) has indicated that category maps are ideal for browsing or simply exploring an information space, but are not good for searching as they don't conform to the common human searching mental models. We have already discovered that cancer researchers have very limited time and therefore do not browse cancer information. We speculate that this tool will be most useful for naive users who are unfamiliar with the cancer information space and/or the terms used in cancer information.

Some of the problems with the category maps have to do with the quality of the labels used to identify the clusters that the algorithm found. We are trying to improve the quality of the labels (in our original document descriptor term identification algorithms) to enhance the quality of the category maps. Pilot studies have indicated that this type of clustering tool is ideal for organizing an information collection into a few general topics and hence can give a good general overview of a given sub-set of a collection, leading us to change our notion of how the category map might be useful in a suite of data mining tools and encouraging us to pursue a dynamic SOM data mining application, which is described in the next section.

4.4 *Information visualization techniques*

– **Graphic Concept Space Representation** – User feedback has indicated that many users are more comfortable with graphical displays of the relationships among document descriptors than with lists of document descriptors (our current method of displaying data mining results). This Java-based tool is simply a graphical representation of the top 10 related terms (10 nearest neighbors) for each input search term. It can be used in conjunction with either a concept space based on our automatic indexing technique or based on the Arizona Noun Phraser (ANP) technique and is similar to the graphical representation that Alta Vista uses. Users can graphically navigate or explore by expanding a term in the tree (creating a sub-tree of that term's 10 nearest neighbors). Figure 8 is an example of the user interface for this tool, which we speculate will be most useful to users who fit either of the concept spaces user profiles but are more graphically oriented.

– **Dynamic SOM** – Initial user feedback also indicated that once a set of documents for a given input term or terms is retrieved, it would be nice to get a big picture of the relationships among those retrieved documents and of their major concepts. This is especially true when the set of documents retrieved is large (over 40 documents) or the topic is new or unfamiliar to the user. To this end, we developed an SOM that creates on command (on the fly) a mini category map for *just* the retrieved documents. Initial feedback has been very positive and we are currently working on tuning and training parameters to optimize the process. Figure 9 is an example of a dynamically created SOM. The user wanted to see the overall landscape of documents that were related to "breast cancer" and "family history".

4.5 *Unified Medical Language System (UMLS) Metathesaurus*

The Unified Medical Language System (UMLS) Metathesaurus was manually created by the National Library of Medicine (NLM), who designers have allowed it to have special word relationships not available to statistically-based techniques. For example, the Metathesaurus contains the following relationships: synonyms, parent terms, children terms (parent and children in an IS-A relationship), broader and narrower terms, and terms related in other ways (similar terms that are not synonyms). Since the UMLS is based on several existing standard medical vocabularies, we believe it will be most useful to medical experts who are familiar with those vocabularies. We have ranked terms according to the above relationships and hope to be able to use UMLS relationships to group terms by level of abstraction (hierarchically) or to assist us in identifying synonymous terms. Figure 10 illustrates the CancerLit UMLS tool.

5. Discussion and Conclusions

Now that government agencies can provide continual public on-line access to vast collections of information through the Internet, those agencies need to make it easier to access that information in meaningful ways. One promising approach is the application of data mining and KDD (knowledge discovery in databases) techniques. The National Cancer Institute is one federal agency that has addressed this challenge by providing public on-line access to several biomedical collections and constantly seeking new technologies and new methodologies to improve accessibility (particularly to CancerLit and PDQ – Physicians Data Query).

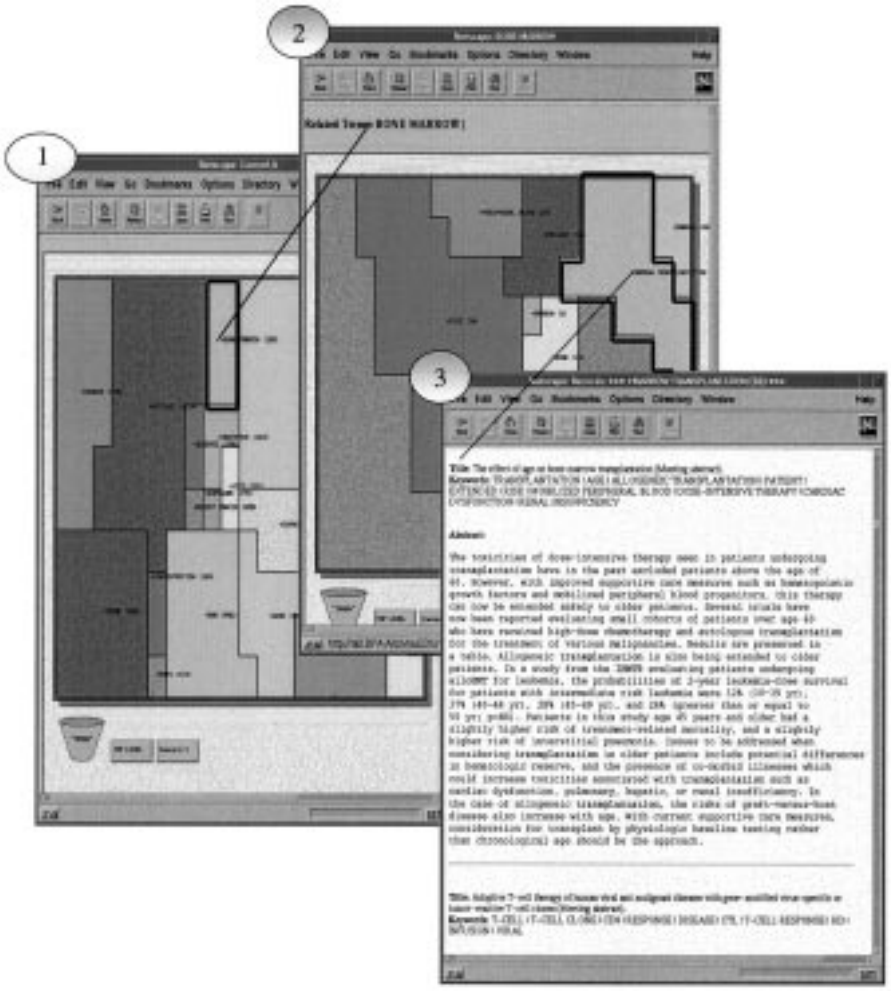

Figure 7. Multiple Layer CancerLit Category Map Session. (1) Shows top layer map. User selects “Bone Marrow” area of map. (2) Shows the 2nd level “Bone Marrow” map. User selects “marrow transplantation” area of map. (3) Shows first document in “marrow transplantation” area.

To quickly and efficiently cope with the huge volumes of information, architectures and techniques that enable intelligent use of this type of data must be scalable. They must be flexible in order to accommodate a diversity of knowledge sources and a variety of data types as well as medical information users who have a wide range of expertise, knowledge source familiarity and information usage requirements. Any system that supports public access to large volumes of information (especially information that has the potential

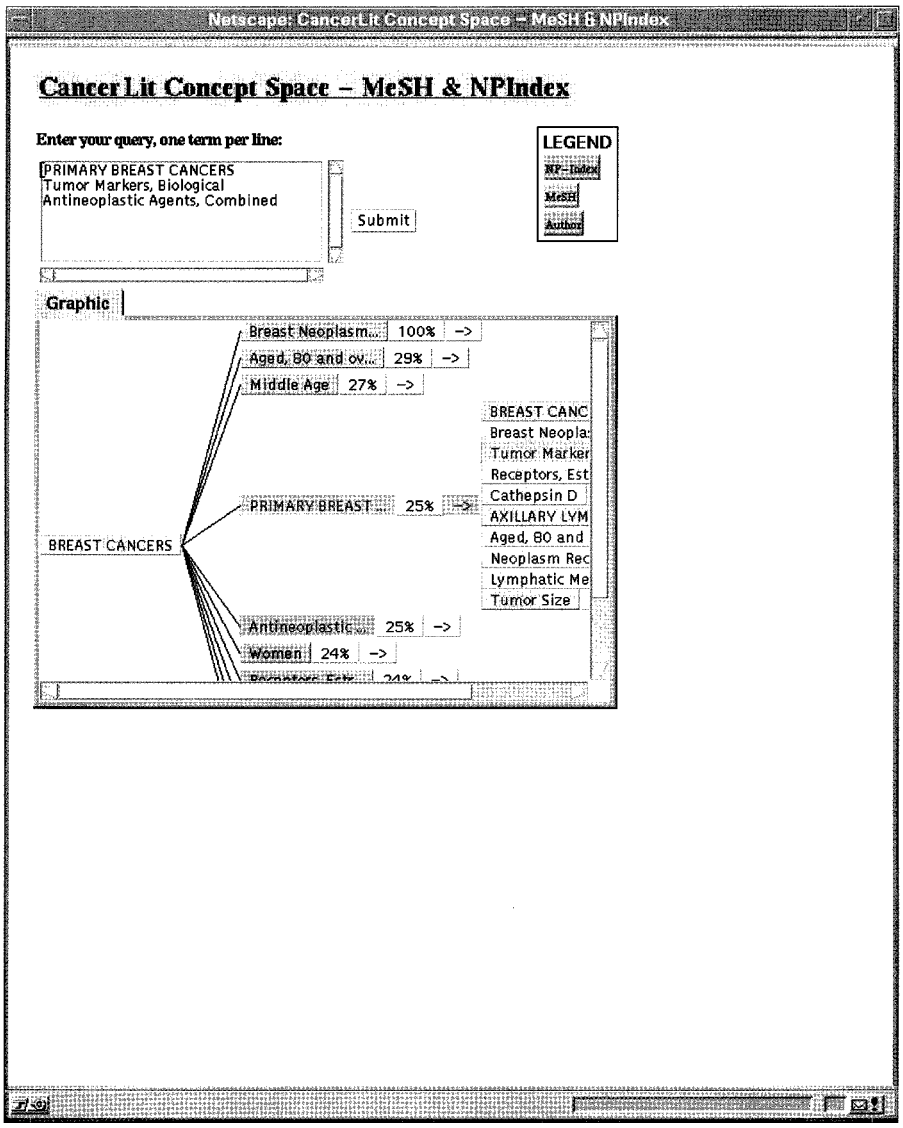

Figure 8. Graphical Concept Space Representation for CancerLit: User selected "breast cancer" as input term.

to be highly technical and complex) must be easy-to-use in terms of its intuitiveness, response time, and availability. Users must be able to identify and retrieve interesting information and relationships in manageable units they can understand, analyze and use. The diversity and amount of information

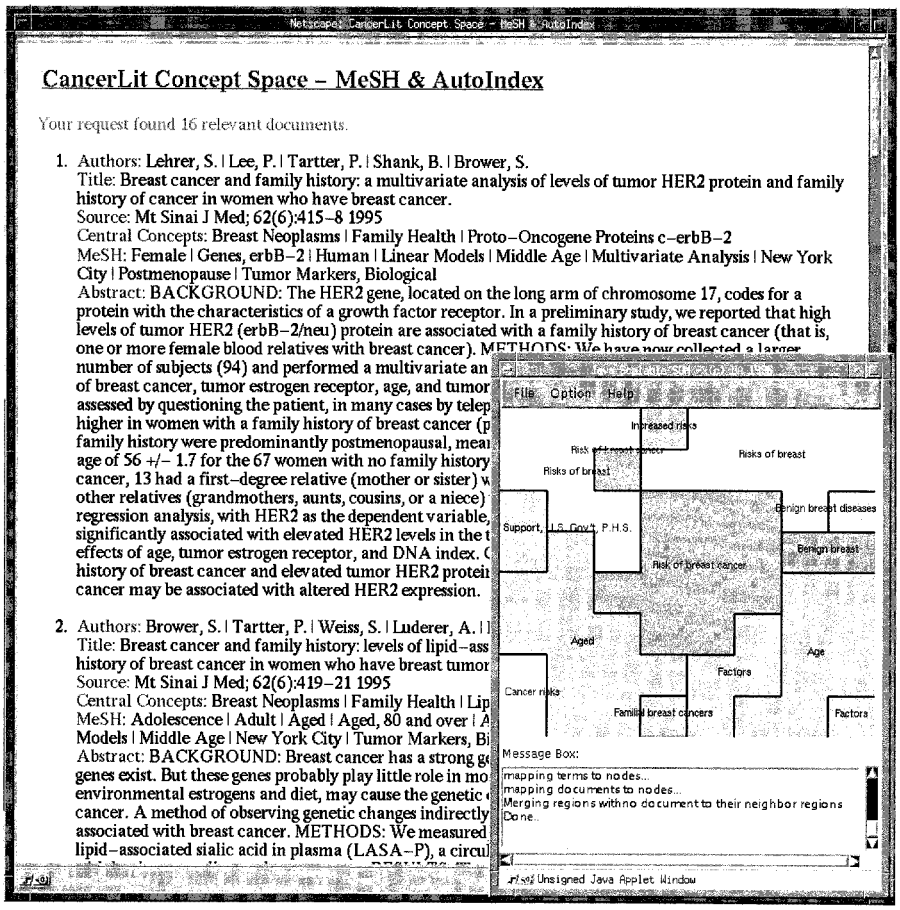

Figure 9. Dynamically Created Document SOM for CancerLit: documents related to "breast cancer" and "family history".

involved suggest that a modular, automated approach, (especially to applied to the clustering types of data mining problems), may be most appropriate.

We have developed and are in the process of testing and refining a medical information retrieval architecture which we feel meets the requirements of this type of information usage. It is comprised of a suite of information data mining tools which we believe can help NCI better manage and improve public access to its extensive cancer information collections.

We have tried to address some of the challenges involved in data mining as follows. First, we have taken advantage of humanly defined classification systems as one way of organizing (classifying or indexing) the documents in the collection. We include MeSH terms, the medical subject headings that are

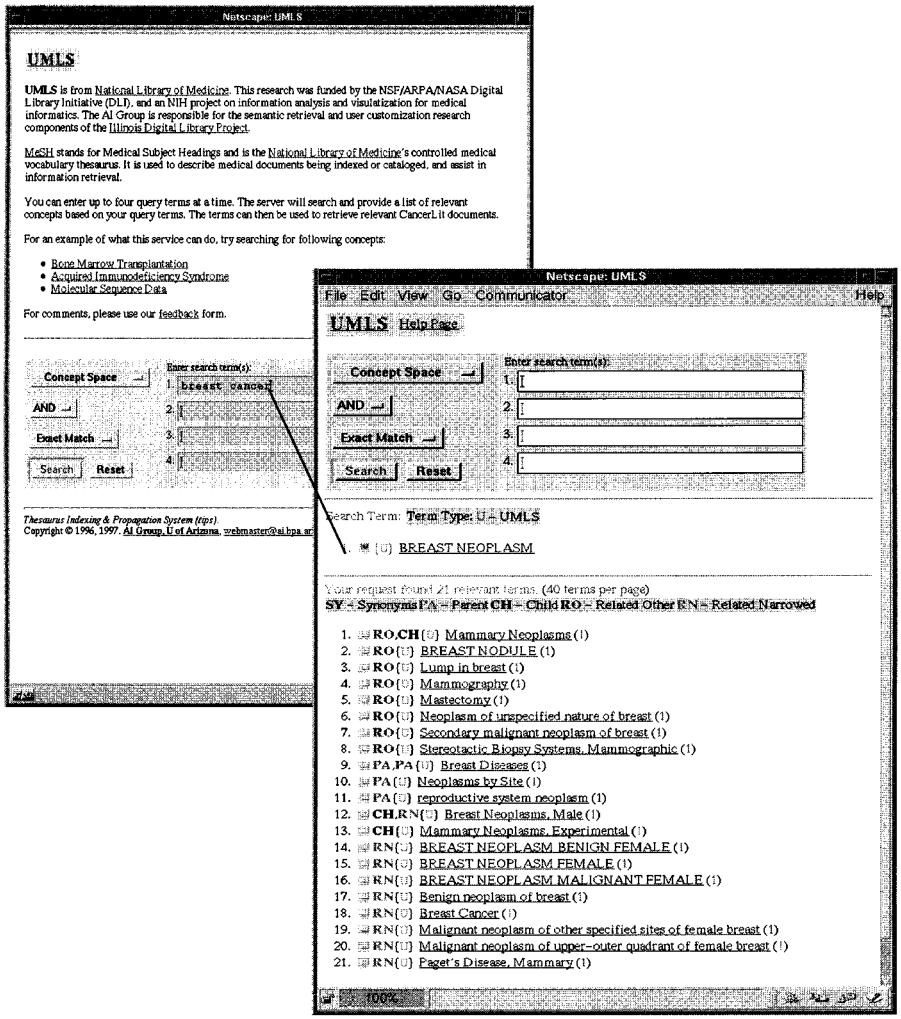

Figure 10. UMLS Session. User inputs "breast cancer". UMLS identifies "Breast cancer" as a synonym for "Breast Neoplasm", the preferred version, and lists all terms identified as related to "Breast Neoplasm".

part of an existing hierarchically organized standard medical indexing vocabulary and have been assigned to each document as its document descriptors by experienced medical indexers at the National Library of Medicine. We also take advantage of syntax, a way of classifying words in sentences or phrases, to help identify a class of document descriptors, namely noun phrases. We believe noun phrases are more accurate than document descriptors identified by term proximity. A series of techniques including a term weighting

algorithm, term frequency, stopwords, and term thresholds are used to reduce the number of document descriptors used in the data mining step. Each of these attempts to reduce the data dimensionality from the almost unlimited number of potential words in a document to some manageable number of important document descriptors.

We have used both a Hopfield and a Kohonen neural net algorithm to cluster data and identify relationships between document descriptors, in part because neural nets deal particularly well with noisy data (Chen et al. 1996a; Chen et al. 1994). Unfortunately, although neural nets can handle multiple dimensions of data and data relationships, the output is typically not easy for humans to understand. We are investigating a variety of visualization techniques to improve the understandability of the output of our data mining tools. Currently we are in the process of expanding our initial pilot studies to find ways of tuning the tools to improve the quality of tool output. One promising future research direction involves allowing the user to identify synonymous terms to reduce the number of obvious or uninteresting relationships between document descriptors.

Other future research directions include exploring ways to improve these techniques through integration with other data mining techniques and the use of new methods of information visualization (for example Multi-Dimensional Scaling, 3-dimensionality, and virtual reality). We are also exploring using these techniques on new knowledge sources. Most of our previous research has been full-text based. We are now moving into the more challenging areas of data mining on image, graphical and tabular data as well as integrated collections.

Acknowledgements

This project was funded primarily by a grant from the National Cancer Institute (NCI) “Information Analysis and Visualization for Cancer Literature” (1996–1997), a grant from National Library of Medicine (NLM), “Semantic Retrieval for Toxicology and Hazardous Substance Databases” (1996–1997), an NSF/CISE “Intelligent Internet Categorization and Search” project (1995–1998), and the NSF/ARPA/NASA Illinois Digital Library Initiative project, “Building the Interspace” (1994–1998).

We would like to thank the following individuals for their generous donation of time and their thoughtful evaluations and suggestions: Dr. Johnathan Silverstein, Dr. Margaret Briel, Dr. Sherry Chow, Dr. Bernie Futcher, Dr. Jesse Martinez, Dr. Luke Whitecell and Nuala Bennett.

We would also like to thank Nick Martin and Dr. Mike Arluk of NCI for their assistance and technical advice in providing the CancerLit collection, and conducting preliminary evaluations of the system.

The CancerLit collection is available from many sources including:

1. CancetLit [database online] Bethesda (MD): National Cancer Institute, 1997. Available through the National Library of Medicine, Bethesda, MD, and
2. CancerLit [Internet URL] http: //www.ovid.com. Available through Ovid Technologies, New York, NY.

Finally, we would like to thank the National Library of Medicine for providing a copy of the Unified Medical Language System (UMLS) to us on CD-ROM and for guidance in our development of our UMLS tool.

References

Agrawal, R., Imielinski, T. & Swami, A. (1993). Database Mining: A Performance Perspective. *IEEE Transactions on Knowledge and Data Engineering* **5**(6): 914–925.

Brill, E. (1993). *A Corpus-based Approach to Language Learning*. PhD thesis, The University of Pennsylvania, Philadelphia, PA.

Chen, H., Chung, Y., Houston, A. L., Li, P. C. & Schatz, B. R. (1999). Using Neural Networks for Vocabulary Switching. Submitted to *IEEE Expert*.

Chen, H., Houston, A., Yen, J. & Nunamaker, J. F. (1996a). Toward Intelligent Meeting Agents. *IEEE COMPUTER* (August) **29**(8): 62–70.

Chen, H., Houston, A. L., Swell, R. R. & Schatz B. R. (1998a). Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques. *Journal of the American Society for Information Science* (May) **49**(7): 582–603.

Chen, H., Hsu, P., Orwig, R., Hoopes, L. & Nunamaker, J. F. (1994). Automatic Concept Classification of Text from Electronic Meetings. *Communications of the ACM* (October) **37**(10): 56–73.

Chen, H. & Lynch, K. J. (1992). Automatic Construction of Networks of Concepts Characterizing Document Databases. *IEEE Transactions on Systems, Man and Cybernetics* (September/October) **22**(5): 885–902.

Chen, H., Lynch, K. J. Basu, K. & Ng, D. T. (1993). Generating, Integrating, and Activating Thesauri for Concept-Based Document Retrieval. *IEEE EXPERT, Special Series on Artificial Intelligence in Text-Based Information Systems* (April) **8**(2): 25–34.

Chen, H. & Ng, D. T. (1995). An Algorithmic Approach to Concept Exploration in a Large Knowledge Network (Automatic Thesaurus Consultation): Symbolic Branch-and-Bound vs. Connectionist Hopfield Net Activation. *Journal of the American Society for Information Science* (June) **46**(5): 348–369.

Chen, H., Schuffels, C. & Orwig, R. (1996b). Internet Categorization and Search: A Machine Learning approach. *Journal of Visual Communications and Image Representations* (March) **7**(1): 88–102.

Chen, H., Zhang, Y. & Houston, A. L. (1998b). Semantic Indexing and Searching Using a Hopfield Net. *Journal of Information Science (JIS)* (January) **24**(1): 3–18.

Dalton, J. & Deshmane A. (1991). Artificial Neural Networks. *IEEE Potentials* (April) **10**(2): 33–36.

Decker, K. M. & Focardi, S. (1995). *Technology Overview: A Report on Data Mining*. Technical Report CSCS TR-95-92, Swiss Scientific Computing Center.

Fayyad, U. M., Piatetsky-Shapiro, G. & Smyth, P. (1996a). From Data Mining to Knowledge Discovery: An Overview. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. & Uthurusamy, R. (eds.) *Advances in Knowledge Discovery and Data Mining*, 1–36. AAAI Press/MIT Press.

Fayyad, U. M., Piatetsky-Shapiro, G. & Smyth, P. (1996b). From Data Mining to Knowledge Discovery in Databases. *AI Magazine* **17**(3): 37–54.

Holsheimer, M. & Siebes, A. P. J. M. (1994). *Data Mining: The Search for Knowledge in Databases*. Technical Report CS-R9406, CWI: Dutch National Research Center.

Honkela, T., Kaski, S., Kohonen, T. & Lagus, K. (1998). Self-Organizing Maps of Very Large Document Collections: Justification for the WEBSOM Method. In Balderjahn, I., Mathar, R. & Schader, M. (eds.) *Classification, Data Analysis, and Data Highways*, 245–252. Springer: Berlin.

Honkela, T., Kaski, S., Lagus, K. & Kohonen, T. (1996a). *Newsgroup Exploration with WEBSOM Method and Browsing Interface*. Techical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland.

Honkela, T., Kaski, S., Lagus, K. & Kohonen, T. (1996b). Self-Organizing Maps of Document Collections. *ALMA* **1**(2). Electronic Journal, address: (http://www.diemme.it /luigi/alma.html).

Hopfield, J. J. (1982). Neural Network and Physical Systems with Collective Computational Abilities. *Proceedings of the National Academy of Science, USA* **79**(4): 2554–2558.

Houston, A. L., Chen, H., Schatz, B. R., Hubbard, S. M., Doszkocs, T. E., Swell, R. R., Tolle, K. M. & Ng, T. D. (1996). *Exploring the Use of Concept Space, Category Map Techniques and Natural Language Parsers to Improve Medical Information Retrieval*. Technical report, University of Arizona, AI Group Working Paper, January.

Houston, A. L., Chen, H., Schatz, B. R., Hubbard, S. M., Swell, R. R. & Ng, T. D. (1998). Exploring the Use of Concept Space to Improve Medical Information Retrieval. *International Journal of Decision Support Systems*, forthcoming.

Hubbard, S. M., Martin, N. B. & Thurn, A. L. (1995). NCI's Cancer Information Systems-Bringing Medical Knowledge to Clinicians. *Oncology* (April) **9**(4): 302–314.

Kashi, S., Honkela, T., Lagus, K. & Kohonen, T. (1996). Creating an Order in Digital Libraries with Self-Organizing Maps. In *Proceedings of WCNN '96, World Congress on Neural Networks, September 15–18, San Diego, California*, 814–817. Mahwah, NJ: Lawrence Erlbaum and INNS Press.

Khosla, R. & Dillon, T. (1997). Knowledge Discovery, Data Mining and Hybrid Systems. In *Engineering Intelligent Hybrid Multi-Agent Systems*, 143–177. Kluwer Academic Publishers.

Knight, K. (1990). Connectionist Ideas and Algorithms. *Communications of the ACM* (November) **33**(11): 59–74.

Kohonen, T. (1989). *Self-Organization and Associate Memory*, 3rd edn. Springer-Verlag: Berlin Heidelberg.

Kohonen, T. (1995). *Self-Organization Maps*. Springer-Verlag: Berlin Heidelberg.

Lippmann, R. P. (1987). An Introduction to Computing with Neural Networks. *IEEE Acoustics Speech and Signal Processing Magazine* (April) **4**(2), 4–22.

Ritter, H. & Kohonen, T. (1989). Self-Organizing Semantic Maps. *Biological Cybernetics* **61**: 241–254.

Salton, G. (1989). *Automatic Text Processing*. Addison-Wesley Publishing Company, Inc.: Reading, MA.

Tank, D. W. & Hopfield, J. J. (1987). Collective Computation in Neuronlike Circuits. *Scientific American* (December) **257**(6): 104–114.

Tolle, K. M. (1997). *Improving Concept Extracting from Text Using Natural Language Processing Noun Phrasing Tools: An Experiment in Medical Information Retrieval*. Master's thesis, Unversity of Arizona, Department of MIS, Tucson, AZ, May.

Uthurusamy, R. (1996). From Data Mining to Knowledge Discovery: Current Challenges and Future Directions. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. & Uthurusamy, R. (eds.) *Advances in Knowledge Discovery and Data Mining*, 561–572. AAAI Press/MIT Press.

Varnado, S. (1995). The Role of Information Technology in Reducing Health Care Cost. In *Proceedings of SPIE – The International Society for Optical Engineering volume 2618: Health Care Information Infrastructure*, 36–46. Philadelphia, PA, October.